The Model is the Message

Isabelle Levent
Stanford University
ilevent@stanford.edu

Lila Shroff
Stanford University
lilashroff@stanford.edu

1 THE MODEL IS THE MESSAGE

A large language model is not simply a tool for text generation—it is a new medium for content, shaping textual artifacts through algorithms and data. Marshall McLuhan famously introduced the idea that “the medium is the message,” the form of content instigates or accelerates cultural shifts [15]. McLuhan presented TV as a medium that revolutionized the political system in the West by creating a new type of national leader who was more of a “showman” than a politician [16]. In this way, LLMs are a new medium that support the creation of machine-generated text. To understand the influence of the form, we need to first understand the process by which text generation occurs. LLMs and, by proxy, the text they generate, are a direct product of data (collection, curation, annotation, etc.), algorithms (encoding and decoding), and computational resources [8, 11, 17]. These components mold different political and technical aspects of the medium from scale to diversity of output to accessibility. Of particular concern to us is how the centralization of LLM development—a product of economic forces around data and compute—shapes writers’ and readers’ control over and experiences of language production and consumption. In the following section, we present historical examples of language control, abuse, and change that situate LLMs within a history of other media mediums and provide a foundation for understanding how LLMs can have deep influence over language.

2 LANGUAGE AND POWER

2.1 Politics of Language Control

In a 1979 essay, author James Baldwin argues against the notion that Black English isn’t a language, writing: “people evolve a language in order to describe and thus control their circumstances” [3]. Historically, social and political forces have asserted authoritative control over language to subvert the agency of less powerful groups. Educational policies forbidding Native American students to learn and speak their native languages played a critical role in forced assimilation and destruction of culture [18]. More recently, deliberate efforts toward language erasure have been cited as fueling the Rohingya genocide in Myanmar [12]. Those who wield power over language are capable of preventing personal expression and erasing group identities.

2.2 Historical Corporate Control Over Language

Control over language can also come through commercial forces: booksellers played a critical role in shaping European nation states. Until the advent of the printing press, there had been no need to standardize language, but when the initial target market for booksellers—literate Latin speakers—grew saturated, vendors began to seek new markets. Europe’s various dialects were then “assembled...within definite limits...into print languages far fewer in number” [1]. National identities formed around these print languages and those dialects which varied most from print language were thrown aside. Notably, this process, unlike earlier examples of social and state abuses of power, was partly arbitrary and had the positive effect of enabling the proliferation of print and information. As political scientist Benedict Anderson writes: “the fixing of print languages...was a largely unselfconscious process resulting from the explosive interaction between capitalism, technology and human linguistic diversity” [1].

2.3 Contemporary Corporate Control Over Language

As LLMs become a mainstream medium for writing, corporations become mediators of everyday language. The immense amount of computational power needed to train LLMs means that the ability to develop new models is restricted to those with an abundance of resources—namely, existing tech giants like Google and Meta, and newer, well-funded players like OpenAI [5]. A report from Europol cites an expert prediction that as much as 90 percent of content online will be synthetically generated by 2026 [7]. While it is impossible to predict the exact ways in which LLMs will come to influence language, we believe it critical to proactively explore concerns of centralized power over language production tools. Such an understanding will allow us to better design and develop tools that meet the human-centered values around language production, rather than defaulting to profit-incentivized commercial decisions.

3 LANGUAGE HOMOGENIZATION

3.1 Centralized Models and The Writing Processes

Existing research on AI-mediated communication gives us a sense of how LLM writing assistants might shape writers’ processes in the future [9]. AI-MC research generally refers to direct communication between people (e.g. over email or text) with AI being present through auto-correct, smart replies, grammar suggestions, and other auto-complete methods [9]. These seemingly innocuous AI suggestions impact language use. Gmail’s “smart reply function” has been found to prime the content of a sender’s response independently of whether or not the sender uses generated suggestions [9]. Biased suggestions for auto-complete in a restaurant review setting skewed a writer’s final review, leading to more positive evaluations [2]. With LLMs now generating large pieces of text and acting as collaborative writing partners, existing concerns about priming and bias are combined with the issue of corporate capture, leading to concerns of language homogenization.

∗Both authors contributed equally to this research.
3.2 Homogenization and the Writer
To expand AI-MC research to questions of language homogenization, we must first define what is meant by language homogenization. The idea of “algorithmic monoculture” refers to multiple actors relying “on the same algorithm” to make decisions and can result in “outcome homogenization” where “particular individuals or groups experience negative outcomes from all decision-makers” [4, 6, 13]. Thus, algorithmic monoculture, in the context of LLMs, involves multiple actors relying on the same model to generate text and can result in language homogenization. Language homogenization may broadly refer to the decrease in syntactic and semantic distance between textual objects that were previously more distinct. More specifically, types of language homogenization may include: (a) homogenization of conversational tones e.g. conversations skew positive online as a result of machine influence (b) homogenization of language used across genres e.g. the language used by a 5th grader writing a final project and a lawyer preparing a brief more closely resemble one another (c) homogenization of language production within an individual’s own corpus e.g. the way one individual writes in personal and work contexts grows more similar. This is a working definition that requires further refinement. More research should be done to identify how significant the threat of language homogenization is, as well as other types of homogenization that may emerge from the centralized development of LLMs. We must also ask the moral and philosophical questions about the contexts in which these acts of homogenization are harmful (e.g. a language model struggling to write a story with a lesbian romance plot without inserting male characters) or helpful (e.g. in aiding a non-native English speaker to produce fluent-appearing text in professional contexts) [10, 14].

3.3 Homogenization and the Reader
Language homogenization is also an issue for the reader who does not know they are encountering machine-generated content. A reader’s approach to evaluating information received from a tweet versus an academic research paper differs due to an acknowledgment of the medium in which the reader is encountering information. Machine-generated content, however, is not visually demarcated by its form like a tweet or a research paper. Instead, this content exists as an invisible layer that may appear to take on a distinct visible form—e.g. machine-generated content masked as a news article—and can be encountered in any context. Without context as to the process by which the media they are consuming has been generated, readers lack the ability to evaluate it genuinely. The role and value of transparency of process in text production must be further evaluated. More research should be done to understand how knowing what model and the extent to which it was used in text generation affects a reader’s evaluation of a text across contexts.

4 CONCLUSION
To have power over language is to hold deep influence over society. Historical examples of corporate power over language show how commercial decisions lead to language change—though often in arbitrary and unexpected ways. We examine LLMs as a new medium and build on the notion of algorithmic monoculture to raise concerns that language production, when mediated by a few centralized corporate players, may lead to language homogenization. We call for further research within language homogenization, including an exploration of the various forms language homogenization may take, an evaluation of when such homogenization is helpful or harmful, and what can be done to mitigate harm in these negative cases.

ACKNOWLEDGMENTS
We would like to thank the following individuals for helping develop our opinions on this topic: Rishi Bommasani, Katy Ilonka Gero, Ziv Epstein, and Michele Elam. We thank Rob Reich for his dedicated mentorship and guidance. We thank Christine Raval, Vanessa Parli, and Stanford Human-Centered Artificial Intelligence for their support. Finally, we thank the members of our HAI-funded student group, Generative AI and the Future of Creative Expression, for their insights and wisdom: Mina Lee, Millie Lin, and Alberto Tono.

REFERENCES
[14] Mina Lee, Megha Srivastava, Amelia Hardy, John Thuckston, Eun Durnus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong,

